Задание 3

1 Описание

Временной ряд — это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.

Стационарность в широком смысле — случайный процесс, у которого математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений ($t_1 - t_2$).

Автокорреляционная функция — зависимость взаимосвязи между функцией (сигналом) и её сдвинутой копией от величины временного сдвига.

Автокорреляционная функция играет важную роль в математическом моделировании и анализе временных рядов, показывая характерные времена для исследуемых процессов.

Для детерминированных сигналов автокорреляционная функция (АКФ) сигнала $f(t)$ определяется интегралом:

$$\Psi(\tau) = \int_{-\infty}^{\infty} f(t) f^*(t-\tau) \mathrm{d}t$$

и показывает связь сигнала (функции $f(t)$ с копией самого себя, смещённого на величину τ . Звёздочка означает комплексное сопряжение.

Для случайных процессов АКФ случайной функции $X(t)$ имеет вид:

$$K(\tau) = \mathbb{E}\{X(t) X^*(t-\tau)\},$$

где $\mathbb{E}\{\ \}$ — математическое ожидание, звёздочка означает комплексное сопряжение.

Простое скользящее среднее, или арифметическое скользящее среднее (англ. simple moving average, англ. SMA) численно равно среднему

арифметическому значений исходной функции за установленный период и вычисляется по формуле:

$$SMA_t = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-i} = \frac{p_t + p_{t-1} + \cdots + p_{t-i} + \cdots + p_{t-n+2} + p_{t-n+1}}{n},$$

где SMA_t — значение простого скользящего среднего в точке t; n - количество значений исходной функции для расчёта скользящего среднего, чем шире сглаживающий интервал, тем более плавным получается график функции; p_{t-i} — значение исходной функции в точке $t - i$

Стандартное отклонение — показывает, на сколько в среднем отклонился ряд от средней вариации ряда (от среднего арифметического, в нашем случае).

Скользящая средняя вместе со стандартным отклонением составляют скользящие статистики.

2 Тест Дики-Фуллера

Тест Дики — Фуллера (DF-тест, Dickey — Fuller test) — это методика, которая используется для анализа временных рядов для проверки на стационарность. Является одним из тестов на единичные корни.

Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд. Это условие записывается как $y_t \sim I(1)$ если ряд первых разностей $\triangle y_t = y_t - y_{t-1}$ является стационарным $\triangle y_t \sim I(0)$.

При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1)

$y_t = a \cdot y_{t-1} + \varepsilon_t$,

где y_t — временной ряд, а ε — ошибка.

Если $a = 1$, то процесс имеет единичный корень, в этом случае ряд y_t не стационарен, является интегрированным временным рядом первого порядка — $I(1)$.

Если $|a| < 1$, то ряд стационарный — $I(0)$.

Для финансово-экономических процессов значение $|a| > 1$ не свойственно, так как в этом случае процесс является «взрывным». Возникновение таких процессов маловероятно.

Приведенное авторегрессионное уравнение AR(1) можно переписать в виде:

$\triangle y_t = b \cdot y_{t-1} + \varepsilon_t$,

где $b = a - 1$, а $\triangle$ — оператор разности первого порядка

$\triangle y_t = y_t - y_{t-1}$.

- Нулевая гипотеза: H_0: $b = 0$ -— процесс нестационарен
- Альтернативаная гипотеза: H_1: $b < 0$ — процесс стационарен

3 Тренд, сезональность, остаток. Аддитивная и мультипликативная модели.

Тренд — тенденция изменения показателей временного ряда. Тренды могут быть описаны различными функциями — линейными, степенными, экспоненциальными и т. д.

Сезонность - периодически колебания, наблюдаемые на временных рядах.

Остаток - разница между предсказанным и наблюдаемым значением.

Аддитивная модель имеет вид: $Y = T + S + E$;

Мультипликативной модель имеет вид: $Y = T * S * E$;

T - тренд, S - сезональность, E - остаток

4 Порядок интегрированности

Интегрированный временной ряд — нестационарный временной ряд, разности некоторого порядка от которого являются стационарным временным рядом. Такие ряды также называют разностно-стационарными (DS-рядами, Difference Stationary).

Для определения интегрированных временных рядов необходимо определить класс временных рядов, называемых стационарными относительно тренда рядами (TS-рядами, trend stationary).

Ряд x_t называется TS-рядом, если существует некоторая детерминированная функция f(t), такая что разность $x_t - f(t)$ является стационарным процессом. В частности, к TS-рядам относятся все стационарные ряды. Однако, многие TS-ряды являются нестационарными.

Временной X_t ряд называется интегрированным порядка k (обычно пишут $X_t \sim I(k)$), если разности ряда k-го порядка $\triangle^k\ x_t$ — являются стационарными, в то время как разности меньшего порядка (включая

нулевого порядка, то есть сам временной ряд) не являются TS-рядами. В частности $I(0)$-это стационарный процесс.

5 Модель ARIMA

ARIMA (англ. autoregressive integrated moving average, иногда модель Бокса — Дженкинса, методология Бокса — Дженкинса) — интегрированная модель авторегрессии — скользящего среднего — модель и методология анализа временных рядов.

Модель $ARIMA(p,d,q)$ означает, что разности временного ряда порядка d подчиняются модели $ARMA(p,q)$.

Модель $ARIMA(p,d,q)$ для нестационарного временного ряда X_t имеет вид:

$$\triangle^d X_t = c + \sum_{i=1}^{p} a_i \triangle^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t,$$

где ε_t — стационарный временной ряд;

c,a_i,b_j — параметры модели. $\triangle^d$ — оператор разности временного ряда порядка d (последовательное взятие d раз разностей первого порядка - сначала от временного ряда, затем от полученных разностей первого порядка, затем от второго порядка и т.д.)

Модель авторегрессии — скользящего среднего (англ. autoregressive moving-average model, ARMA) — одна из математических моделей, использующихся для анализа и прогнозирования стационарных временных рядов в статистике. Модель ARMA обобщает две более простые модели временных рядов — модель авторегрессии (AR) и модель скользящего среднего (MA).

Моделью $ARMA(p, q)$, где p и q — целые числа, задающие порядок модели, называется следующий процесс генерации временного ряда X_t:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i},$$

где c — константа, ε_t — белый шум, то есть последовательность независимых и одинаково распределённых случайных величин (как правило, нормальных), с нулевым средним, а $\alpha_1, \ldots, \alpha_p$ и $\beta_1, \ldots, \beta_q$ — действительные числа, авторегрессионные коэффициенты и коэффициенты скользящего среднего, соответственно.

Авторегрессионная (AR-) модель (англ. autoregressive model) — модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка p (AR(p)-процесс) определяется следующим образом

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t,$$

где $a_1, \ldots, a_p$ — параметры модели (коэффициенты авторегрессии), c — постоянная (часто для упрощения предполагается равной нулю), а ε_t — белый шум.

Модель скользящего среднего q-го порядка $MA(q)$ — модель временного ряда вида:

$$X_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j},$$

где ε_t — белый шум, b_j — параметры модели b_0 можно считать равным 1 без ограничения общности).

$$ARIMA(p, d, q) = AR(p) + MA(q) \sim I(d)$$

6 R2 Score

Коэффициент детерминации R^2 — R-квадрат) — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными.

Истинный коэффициент детерминации модели зависимости случайной величины y от факторов x определяется следующим образом:

$$R^2 = 1 - \frac{V(y|x)}{V(y)} = 1 - \frac{\sigma^2}{\sigma_y^2},$$

где $V(y|x) = \sigma^2$ — условная (по факторам x) дисперсия зависимой переменной (дисперсия случайной ошибки модели).

В данном определении используются истинные параметры, характеризующие распределение случайных величин.

Если использовать выборочную оценку значений соответствующих дисперсий, то получим формулу для выборочного коэффициента детерминации (который обычно и подразумевается под коэффициентом детерминации):

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{SS_{res}/n}{SS_{tot}/n} = 1 - \frac{SS_{res}}{SS_{tot}},$$

где $SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ — сумма квадратов остатков регрессии, y_i,$\hat{y}_i$ — фактические и расчётные значения объясняемой переменной.

$SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = n\hat{\sigma}_y^2$ — общая сумма квадратов.

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость.

При оценке регрессионных моделей это интерпретируется как соответствие модели данным.

Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50% (в этом случае коэффициент множественной корреляции превышает по модулю 70%).

Модели с коэффициентом детерминации выше 80% можно признать достаточно хорошими (коэффициент корреляции превышает 90%). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

7 AIC - информационный критерий Акаике

Информационный критерий — применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень ”подгонки”модели под данные с корректировкой на используемое количество оцениваемых параметров. Т.е. критерии основаны на некотором компромиссе между точностью и сложностью модели.

Критерии различаются тем, как они обеспечивают этот баланс. информационные модели используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих

критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

AIC (an information criterion) — информационный критерий Акаике - критерий, применяющийся исключительно для выбора из нескольких статистических моделей. Связан с концепцией информационной энтропии и расстоянии Кульбака-Лейблера, на основе которой был разработан этот критерий.

В общем случае AIC:

$$AIC = 2k - 2\ln(L),$$

где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели.

Предположим, что ошибки модели нормально и независимо распределены. Пусть n — число наблюдений и RSS-

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

остаточная сумма квадратов. Далее мы предполагаем, что дисперсия ошибок модели неизвестна, но одинакова для всех них. Следовательно:

$$AIC = 2k + n[\ln(2\pi RSS/n) + 1]$$

В случае сравнения моделей на выборках одинаковой длины, выражение можно упростить, выкидывая члены зависящие только от n:

$$AIC = 2k + n[\ln(RSS)]$$

8 Реализация

1. Чтение данных из training.xlsx
   - pd.read_excel(“training.xlsx”, index_col=0)
     Библиотека pandas
2. Проверка на стационарность
   - Визуальная оценка
     - ts.rolling().mean()
       * Возвращает скользящее среднее

        * Библиотека pandas
    - ts.rolling().std()
        * Возвращает стандартное отклонение
        * Библиотека pandas
    - Строим вывод о стационарности ряда по визуальной оценке
- Тест Дики-Фуллера
    - sm.tsa.adfuller(ts)
        * проводит тест Дики-Фуллера
        * ts - временной ряд
        * Возвращает массив с данными => adf, critical values, …
        * Библиотека statsmodels
        * Анализируем полученные из функции параметры
        * Строим вывод о стационарности данного ряда

3. Разложение временного ряда на тренд, сезональность остаток в соответствии с аддитивной и мультипликативной моделями

    - seasonal_decompose(data.Value, model)
        - data.Value - столбец значений временного ряда
        - model - ’additive’ или ’multiplicate’
        - Возвращает decompose
            * decompose.trend - тренд исходного ряда
            * decompose.resid - остаток исходного ряда
            * decompose.seasonal - сезонность исходного ряда
        - Библиотека statsmodels
    - Строим вывод на основании полученных результатов

4. Поиск индекса интегрируемости ряда

    - findk(dat, n)
        - dat.diff(periods=n).
            * Вычисляет разницу между исхожным рядом и рядом со смещением n
        - dropna()
            * Удаляем первые NaN элементы
        - sm.tsa.adfuller(dat)

    * С помощью теста Дики-Фуллера смотрим является ли новый ряд стационарным

5. Подбираем нужные параметры с помощью функции автокорреляции и функции частичной автокорреляции

    - acf(ts, nlags) - находит автокорреляцию временного ряда
        - ts - временной ряд
        - nlags - число лагов для автокорреляции
        - Помогает найти порядок q модели MA(q) для построения модели ARIMA(p, d, q)
        - Возвращает массив с значениями функции
        - Библиотека statsmodels
    - pacf(ts, nlags) - находит частичную автокорреляцию временного ряда
        - ts - временной ряд
        - nlags - число лагов для автокорреляции
        - Помогает найти порядок p модели AR(p) для построения модели ARIMA(p, d, q)
        - Возвращает массив с значениями функции
        - Библиотека statsmodels
    - Функции автокорреляции и частичной автокорреляции - дискретные, число различных значений равно лагу (в обоих случаях по 50).

      p - номер последнего лага, который не входит в доверительный интервал на графике pacf.

      q - номер последнего лага, который не входит в доверительный интервал на графике acf.

6. ARIMA-модель

    - ARIMA(ts, order=(p, k, q)) - строит модель ARIMA на основе временного ряда ts
        - ts - временной ряд
        - order - порядки для модели
        - Возвращает объект ARIMA
        - Библиотека statsmodels

- model.fit() - Прогоняем ARIMA-модель через фильтр Калмана
  - Фильтр Калмана - мощнейщий инструмент фильтрации данных. При фильтрации используется информация о физике самого явления.
  - Возвращает объект ARIMAResults - содержит также и результаты предсказаний, которые можно выводить и использовать в дальнейших прогнозах
  - Метод класса ARIMA
- model.predict(start='1989-01-01', end='1993-12-01', typ='levels')
  - Строим предсказание на тестовой выборке
  - start - определяет с какой позиции начать предсказывать
  - end - определяет до какой позиции предсказывать
  - Возвращает предсказанные значения
  - Библиотека statsmodels

7. R2 Score

- r2_score(y_true, y_pred)
  - подсчитывает r2 score между реальной выборкой и предсказанной
  - y_true - реальные значения
  - y_pred - предсказанные значения
  - Библиотека sklearn.metrics
  - Вовзращает r2 score

8. AIC

- model.aic - выводит значение критерия - вычисляется в процессе построения статистической модели

9 Библиотеки

1. Pandas
    - Для работы с временными рядами
2. Numpy
    - Для работы с массивами
3. Matplotlib.pylab
    - Для построения графиков функций
4. Statsmodels
    - Используются статистические функции
    - Для постороения модели ARIMA
5. Sklearn.metrics
    - Для метрики R2 score